Top-down free-energy minimization on protein potential energy landscapes

Bruce W. Church and David Shalloway*

Biophysics Program, Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853

Edited by R. Stephen Berry, University of Chicago, Chicago, IL, and approved March 21, 2001 (received for review January 18, 2001)

The hierarchical properties of potential energy landscapes have been used to gain insight into thermodynamic and kinetic properties of protein ensembles. It also may be possible to use them to direct computational searches for thermodynamically stable macroscopic states, i.e., computational protein folding. To this end, we have developed a top-down search procedure in which conformation space is recursively dissected according to the intrinsic hierarchical structure of a landscape's effective-energy barriers. This procedure generates an inverted tree similar to the disconnectivity graphs generated by local minima-clustering methods, but it fundamentally differs in the manner in which the portion of the tree that is to be computationally explored is selected. A key ingredient is a branch-selection algorithm that takes advantage of statistically predictive properties of the landscape to guide searches down the tree branches that are most likely to lead to the physically relevant macroscopic states. Using the computational folding of a β-hairpin-forming peptide as an example, we show that such predictive properties indeed exist and can be used for structure prediction by free-energy global minimization.

New methods have been developed in recent years for using the global properties of protein potential energy landscapes to analyze overall thermodynamic and kinetic properties of protein ensembles. The method pioneered by Bryngelson and Wolynes (1) uses order parameters to characterize global dynamic properties such as funneling (2–8). Other methods generate hierarchical inverted trees or disconnectivity graphs, whose topologies reflect selected aspects of landscape structure, by finding and hierarchically grouping local minima according to metrics such as Euclidean distance in conformation space (9–13), potential energy barrier height (14–23), or effective-energy barrier (peak of a potential-of-mean-force) height (24–28). Different methods for finding local minima have been used for this purpose, such as steepest descent quenching from molecular dynamics simulations (10, 11, 14, 15), and the threshold method, which progressively extends a search of the basin surrounding a known local minimum until a new minimum of lower energy is found (21–23). By grouping the catchment regions [steepest-descent minimization starting from any point in a catchment region leads to its local minimum (29)] of the local minima, a disconnectivity graph defines a variable-scale decomposition of conformation space: The ''root'' (the top) is the complete region containing all conformation space, the ''leaves'' (the bottom) are the local minima catchment regions, and each branch at a given level parameter corresponds to an extended region that is connected by virtue of satisfying a level inequality by using the graph's metric.

The branches of trees constructed with the effective-energy metric (24–28), which are parameterized by thermal energy or temperature T (in units where k_B is unity), correspond to the macroscopic thermodynamic states (macrostates) of the system—i.e., to conformation space regions that kinetically confine the system at T (28, 30, 31). (If entropic effects are not too important, the macrostates also can be approximated as branches of trees constructed by using the potential barrier metric.) The top is reached when T exceeds all effective-energy barriers, the leaves (local minima) are at $T = 0$ and the level of experimental relevance is (for biological problems) $T = T_{phys} \sim 310\ ^\circ K \sim 0.6$ kcal/mol.[†] The top macrostate contains the entire space, and macrostates decrease in size and increase in number with decreasing T until they correspond to local minima catchment regions at low T.

It also may be possible to use such hierarchical trees for another purpose—to efficiently guide searches for the macrostates of lowest free-energy (the physically relevant subset of macrostates, PRSM) at T_{phys}; i.e. for global free-energy minimization. Our focus here is the possibility of developing a top-down tree-search method for global minimization that progressively subdivides conformation space and explores only selected branches as T is lowered toward T_{phys}. Although this type of search implicitly relies on the existence of an underlying tree structure, most of the tree is never computed: the utility of a top-down method lies in its ability to find a global minimum while computing very little of the tree. This requires a branch selection algorithm that can choose, during the annealing process itself, the branches that are most likely to lead to low free-energy macrostates at T_{phys}. Branches that are not selected are not explored, resulting in computational efficiency, but at the risk of excluding important regions from subsequent search.

Local minima-based methods are not top-down in this context because they all start from local minima, progress by finding more local minima, and build the higher (and larger) branches of a disconnectivity graph by aggregating local minima (see ref. 21 for an example of a global minimization method that moves from local minima to local minima using the ''threshold'' algorithm and ref. 20 for a method of building a disconnectivity graph starting with knowledge of the global minimum). These methods do not attempt to select the most promising branches by using a branch-selection algorithm. In contrast, a top-down search to nonzero T_{phys} does not seek any local minima. Even if a top-down search were used to find local minima by continuing the search down to $T = 0$, only a few local minima would be computed at the very end of the search process.) Instead, it subdivides branches by using distance-geometry inequalities. Most importantly, this procedure provides a context in which a branch selection algorithm can select the most promising descendent branches for further search.

This paper was submitted directly (Track II) to the PNAS office.

Abbreviation: PRSM, physically relevant subset of macrostates.

*To whom reprint requests should be addressed. E-mail: dis2@cornell.edu.

[†]Within the subclass of methods that use an energetic or temperature metric, three different measures have been used: (*i*) $T = \Delta E/k_B$, where ΔE is the transition state energy barrier (14–23), (*ii*) the T at which the transition rate equals a specified value (28), and (*iii*) the T at which the transition rate equals a specified fraction of the macrostate relaxation rate (24–26). They all generate roughly similar trees, but differ in detail; each has advantages depending on the intended application. Any one could be used for global minimization, but we find the third to be most convenient. In distinction, the best hierarchy for analyzing dynamics by a master equation would be one that used inter-macrostate transition rates at T_{phys} as a metric. Because of the complicated nonlinear relationship between transition rates and temperature [when the temperature dependence of $\Delta E(T)$ and $\Delta S(T)$ are considered], this method may differ in detail from the hierarchies studied to date.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

www.pnas.org/cgi/doi/10.1073/pnas.101030498

A top-down search requires: (*i*) an algorithm for recursive subdivision of macrostates/branches, and (*ii*) an effective branch selection algorithm. We previously have described a method for recursive subdivision based on the identification of effective-energy barriers during a computational cooling process that starts from the top macrostate (25, 26). Stochastic sampling of the landscape is algorithmically monitored during cooling for the appearance of effective-energy barriers that would trap sampling in subregions (i.e., break computational ergodicity), and thus reduce sampling efficiency, if temperature were further lowered (see ref. 32 for an interesting discussion of this problem). Before T is lowered to a point where trapping occurs, the parent macrostate is subdivided into child macrostates that are separated by the effective-energy barrier. Independent search processes can then be spawned (e.g., a coarse-grained parallel computer) for each child that do not need to cross the barrier. This maintains computational ergodicity and efficient sampling at all T at the expense of an increased number of separate search processes.

The second task, effective branch selection, will only be possible if there are inheritance properties of the macrostates at $T > T_{\text{phys}}$ that can be used to partially predict that their descendents will be PRSM members. This is not guaranteed, and it is easy to construct landscapes whose hierarchies have no predictive power (33). Yet it seems plausible that predictive properties exist and can be used to hierarchically solve the global minimization problem with reduced computational cost. Here we empirically explore this key question: Using a β-hairpin-forming octapeptide as an example, we show that this hypothesis is true and compare the utility of different branch-selection algorithms. We conclude that hierarchical top-down searching can be a valuable tool in computational structure prediction.

Methods

We fix bond lengths and angles and sample conformation space by using the protein backbone and side-chain torsion angles, denoted Ω, as variables.

Recursive Computation of Window Functions in Distance Variables. The fundamentals already have been described (24–26, 30); we summarize here: Each macrostate α is specified by a macrostate window function $w_\alpha(T; \Omega)$, which is ~1 within the macrostate and ~0 outside. Here it is adequate to use "hard" window functions that are either 1 or 0, and to assume that the window functions only change discontinuously (i.e., by subdivision) and are otherwise constant between bifurcation temperatures. The top macrostate, 0, includes all conformation space, so $w_0 = 1$. At lower T there are multiple macrostates $\{\alpha\}_T$, whose window functions satisfy $\Sigma_\alpha w_\alpha(T; \Omega) = 1, \forall\Omega$. That is, they completely dissect conformation space.

As T is lowered through a descending sequence of temperatures $\{T_i\}$, each of the macrostates of interest are separately tested for bifurcation (see below). When a macrostate α divides into children β and γ, w_α is divided into w_β and w_γ:

$$w_\beta(\Omega) = \Theta^{\beta\alpha}(\Omega) w_\alpha(\Omega), \qquad [1]$$

$$w_\gamma(\Omega) = \Theta^{\gamma\alpha}(\Omega) w_\alpha(\Omega), \qquad [2]$$

where $\Theta^{\beta\alpha}(\Omega)$ and $\Theta^{\gamma\alpha}(\Omega)$ equal 0 or 1, and $\Theta^{\beta\alpha}(\Omega) + \Theta^{\gamma\alpha}(\Omega) = 1$. Recursive application of Eqs. **1** and **2** yields window functions of the form

$$w_\delta(T; \Omega) = \Theta^{\delta\delta_1}(\Omega)\Theta^{\delta_1\delta_2}(\Omega) \ldots \Theta^{\delta_N 0}(\Omega), \qquad [3]$$

where δ_1 is the parent of δ, δ_2 is the parent of δ_1, and so on up to δ_N, which is a child of the top macrostate.

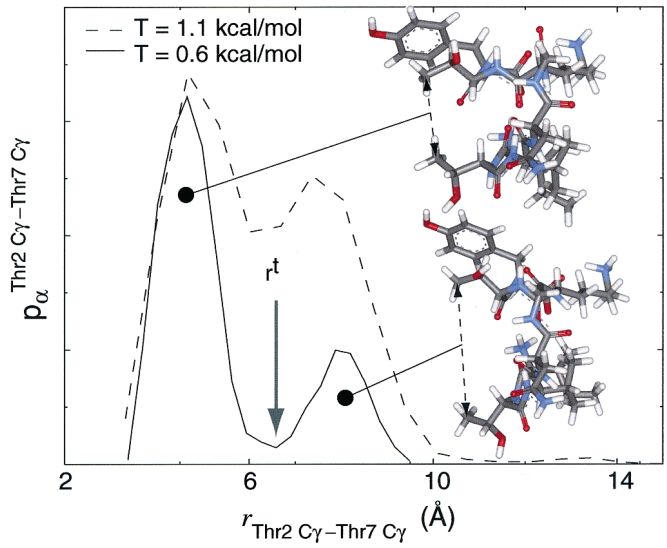

Fig. 1. Probability histograms for BH8 at high and low *T*. The mean structures of the two children after bifurcation are shown; the double-headed arrow identifies the bifurcating distance.

Detecting Bifurcations. The equilibrium probability distribution within macrostate α is

$$p_\alpha(T; \Omega) \propto e^{-V[R(\Omega)]/T} w_\alpha(\Omega),$$

where $V(R)$ is the potential in Cartesian coordinates R. [The Jacobian of the change of variables from R to Ω is ignored because it is independent of Ω and factors out of all conformation space integrations (34).] To detect bifurcations of w_α we first approximate $p_\alpha(T; \Omega)$ as a sum of localized distributions whose number and character (i.e., zero, first and second moments) are determined by characteristic packet equations as described in refs. 24 and 30. It is then simple to define window functions that separate these regions (24, 30). However, applying the packet equations in multidimensional form can be costly. Instead we apply them to one-dimensional effective-energy reaction coordinates that are derived by separately projecting $p_\alpha(T; \Omega)$ onto each of the interatomic distance variables $r_{ij}(\Omega) \equiv |\vec{r}_i(\Omega) - \vec{r}_j(\Omega)|$, where $\vec{r}_i$ is the 3-vector Cartesian coordinate of atom i. Fig. 1 displays one such projected probability distribution, $p_\alpha^{\text{Thr2Cγ–Thr7Cγ}}(\text{T; r})$, at two different temperatures. At $T = 1.1$ kcal/mol the packet equations have only a single solution, so there is only one macrostate, which contains the entire region. When T is lowered to ~0.6 kcal/mol, $p_\alpha^{\text{Thr2Cγ–Thr7Cγ}}(T; r)$ becomes sufficiently bimodal so that two independent child solutions appear corresponding to the two separate concentrations of probability. This signals the bifurcation of the macrostate into two children. The window functions are then defined by Eqs. **1** and **2** with the approximation

$$\Theta^{\beta\alpha}(\Omega) \approx \theta[r_{ij}(\Omega) - r_{ij}^t], \qquad [4]$$

$$\Theta^{\gamma\alpha}(\Omega) \approx \theta[r_{ij}^t - r_{ij}(\Omega)], \qquad [5]$$

where r_{ij}^t is the value of r_{ij} corresponding to the node of p_α^{ij} (see Fig. 1) and θ is the Heaviside step function.

Because the distance variables are highly redundant, we expect that this procedure will identify each macrostate as an isolated concentration of probability in at least one projected representation. In effect, the distance variables provide a large set of possible reaction coordinates that can be examined for confining effective-energy barriers. No artifactual barriers will be introduced by this approximation, although it is possible (though

BIOPHYSICS

unlikely) that an effective-energy barrier could be missed. The danger of missing an effective-energy barrier is that computational ergodicity may be broken within the (spuriously undivided) macrostate. We have not yet encountered any case in which this has occurred.

The use of the r_{ij} as reaction coordinates for analyzing metastability implicitly assumes that probability equilibrates rapidly in the transverse directions. This will not be true for all r_{ij}, but should be true when there is an effective-energy barrier in the coordinate. Because this is the only case in which the *ij* projection will be used, the assumption is self-consistent. To maintain dynamical significance, the projection must account for the fact that the r_{ij} are nonlinear functions of Ω. This is described in ref. 26.

We sampled $p_\alpha(T; \Omega)$ within each macrostate α by using the Metropolis Monte Carlo method with an anisotropic multivariate wrapped Gaussian transition function (35, 36) and the J-walking algorithm (37). Distance-variable probability histograms describing $p^{ij}_\alpha(T; r)$ were computed (typically with 20 bins) for all the interatomic pairs using every tenth sample point. The fractional energy-fluctuation autocorrelation between sample points decayed to ~0.5 after 50 steps. Computing these histograms added little cost compared with the cost of evaluating $V[R(\Omega)]$, because the interatomic distances were already required to compute the potential. The characteristic packet equations were solved by using trapezoidal integration over the histogram bins.

Computing Macrostate Thermodynamic Properties. Intensive properties, such as mean energy,

$$E_\alpha(T) = \frac{\int V[R(\Omega)] p_\alpha(T; \Omega) d\Omega}{\int p_\alpha(T, \Omega) d\Omega}, \qquad [6]$$

were computed from the Metropolis Monte Carlo sampling at each $T = T_i$.

Macrostate entropy $S_\alpha(T)$ is extensive and can not be computed in this manner. Instead, we computed it and macrostate free energy $F_\alpha(T)$ as follows: We fixed the (classical) arbitrary entropy scale by setting $F_0(T_{hi}) = 0$ (i.e., for the top macrostate). When it bifurcated at temperature $T^{\beta\gamma}$ the free energies of its children, β and γ, were calculated from their probability ratio p_β/p_γ. In accord with Eqs. **4** and **5**, this is

$$\frac{p_\beta(T^{\beta\gamma})}{p_\gamma(T^{\beta\gamma})} = \frac{\int p^{ij}_\alpha(T^{\beta\gamma}; r)\theta(r - r^t_{ij}) dr}{\int p^{ij}_\alpha(T^{\beta\gamma}; r)\theta(r^t_{ij} - r) dr}, \qquad [7]$$

where r_{ij} is the distance variable in which the bifurcation occurs. Then, F_β and F_γ at $T^{\beta\gamma}$ were calculated by using

$$\frac{p_\beta(T^{\beta\gamma})}{p_\gamma(T^{\beta\gamma})} = e^{-(F_\beta - F_\gamma)/T^{\beta\gamma}} \qquad [8]$$

and the conservation of probability relationship

$$e^{-F_0/T^{\beta\gamma}} = e^{-F_\beta/T^{\beta\gamma}} + e^{-F_\gamma/T^{\beta\gamma}}. \qquad [9]$$

S_β and S_γ were then calculated by using the thermodynamic relationship

$$F_\alpha = E_\alpha - T^{\beta\gamma} S_\alpha. \qquad [10]$$

As T decreased, the mean energies were updated by using Eq. **6**. Entropies were updated by (discretely) integrating

$$\frac{dS_\alpha(T)}{dT} = \frac{dE_\alpha(T)/dT}{T}. \qquad [11]$$

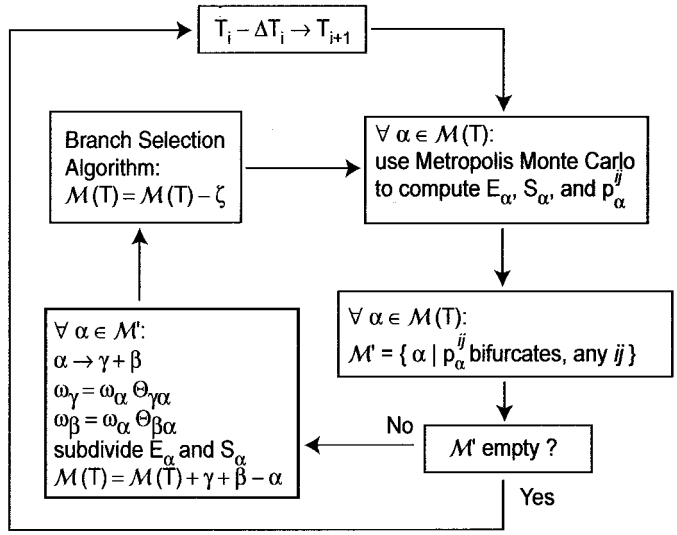

Fig. 2. Algorithm. $\mathcal{M}(T)$ is the set of macrostates that are being tracked at temperature T; α refers to a macrostate and β and γ to its children. $\mathcal{M}(T)$ is first expanded by identifying the subset of macrostates that undergo bifurcation, $\mathcal{M}'$, and recursively bifurcating them and their children until no bifurcations remain. Bifurcations expand $\mathcal{M}(T)$; the branch-selection algorithm is used to prune macrostates ζ before T is lowered according to the cooling schedule.

The derivative of E_α in Eq. **11** was calculated by finite difference. While the derivative could be calculated from the potential energy variance, it is more accurate to use the finite difference because of its faster convergence. Moreover, when the finite difference is used, computational errors in E_α at different T largely cancel when Eq. **11** is integrated.

Annealing and Branch Selection. The algorithm is summarized in Fig. 2. The cooling schedule was empirically chosen to be slow enough so that each macrostate bifurcated before sampling ergodicity was broken:

$$\Delta T_i = T_i - T_{i+1} = \begin{cases} 1.0, & T_i > 4.0 \\ 0.1, & 4.0 < T_i < 1.1. \\ 0.05, & T_i < 1.1 \end{cases} \qquad [12]$$

Each member of the set of macrostates being tracked, $\mathcal{M}(T)$, was sampled and tested for bifurcations at each T_i as described above by using 128,000 sample points, a value that gave $\Delta E_\alpha(T)/T \sim 0.05$ for all macrostates (where ΔE_α is the standard error of the mean.) To handle multifurcations, after each bifurcation, the children were resampled and tested for additional bifurcations at the same temperature. The thermodynamic parameters of the children were computed by partitioning the parental sample points. The parent was replaced in $\mathcal{M}(T)$ by its children. The branch-selection algorithm then was applied to reduce the number of macrostates in $\mathcal{M}(T)$ according to the specified criterion.

Number of Contacts. We defined the number of nonlocal contacts $N_{C,\alpha}(T)$ for macrostate α as the number of r_{ij} for pairs separated by more than one torsion angle that had

$$\int_0^{1.2\, r^{vdW}_{ij}} p^{ij}_\alpha(T; r) dr > 0.5, \qquad [13]$$

where r^{vdW}_{ij} is the van der Waals contact distance for pair ij. The probability-weighted number of contacts

$$\langle N_C\rangle(T) = \frac{\Sigma_\alpha N_{C,\alpha}(T) e^{-F_\alpha/T}}{\Sigma_\alpha e^{-F_\alpha/T}}, \qquad [14]$$

is a T-dependent estimator of overall compactness.

Estimated Number of Macrostates. Not all macrostates were computed, but their T-dependent total number $N_M(T)$ was estimated (assuming that the rate, in T, of bifurcation is similar for the unobserved and observed branches) by calculating the observed average rate of bifurcation, $g(T) \equiv d \log(N_M^{obs})/dT$, and integrating

$$\frac{dN_M(T)}{dT} = N_M(T)g(T), \qquad [15]$$

using the boundary condition $N_M(T_{hi}) = 1$.

Computation. Tree analysis was performed with coarse-grained parallelization on a cluster of Pentium processors using a master-slave configuration. Slaves computed individual macrostate branches independently as scheduled by the master according to the branch-selection algorithm. Therefore, interprocessor communication was minimal, and parallelization efficiency was essentially 100%.

Results and Discussion

Top-Down Discovery of the BH8 Macrostate Tree. To provide a model for examining different branch-selection algorithms, an extensive macrostate tree was generated by using the top-down method for the BH8 octapeptide (ITVNGKTY), a peptide designed to fold into a β-hairpin (38). We used the ECEPP/3 potential (39), an all-atom potential with fixed bond lengths and bond angles, in torsion-angle coordinates, Ω, augmented by empirical solvation based on solvent-accessible surface area (40). As a benchmark for subsequent analysis, a large number of macrostates (all those having equilibrium probability $\geq 10^{-3}$) were computed, even though most would not be computed in an actual free-energy global minimization run guided by a branch-selection algorithm.

The Gibbs–Boltzmann distribution was sampled by using a modified Metropolis algorithm (see *Methods*), starting at a temperature (T_{hi} = 35 kcal/mol) well above the maximum ECEPP/3 barriers to individual torsion-angle rotation. Macrostate thermodynamic properties were computed as T was gradually lowered (see Eq. **12**). In addition, the effective-energy functions that resulted from projecting the Gibbs–Boltzmann probability distribution onto each of the interatomic-distance variables were algorithmically monitored for signals of macrostate bifurcation: When a bifurcation temperature was reached at which an effective-energy barrier appeared that satisfied the bifurcation conditions (which imply macrostate metastability), the macrostate was subdivided by a distance-variable inequality (see *Methods* for details). By this means, large effective-energy barriers were detected during sampling at high T, and small barriers were detected at lower T.

For example, Fig. 1 shows the projected probability histograms for $r_{Thr2C_\gamma - Thr7C_\gamma}$, the distance between the C_γ atoms of Thr-2 and Thr-7 for one macrostate at two different temperatures. The histogram is effectively unimodal at T = 1.1 kcal/mol (i.e., the small dip in probability between the modes does not restrict the transition rate), but is sufficiently bimodal at its bifurcation temperature, 0.6 kcal/mol, to satisfy the bifurcation conditions. Thus, at this temperature the parent macrostate α was subdivided, using $r_{Thr2C_\gamma - Thr7C_\gamma}$ as a reaction coordinate, into children β and γ centered around $r_{Thr2C_\gamma - Thr7C_\gamma}$ ~4.5 Å and $r_{Thr2C_\gamma - Thr7C_\gamma}$ ~8 Å.

Because of the partial redundancy between different distance variables [resulting from the fact that there are $O(N^2)$ distance variables but only $O(N)$ degrees of freedom, where N is the number of atoms], a single effective-energy barrier (i.e., having a unique location in the internal coordinate or Cartesian space) often will manifest as probability gaps in multiple distance variables. When this happens, the bifurcating reaction coordinate will be the one that first satisfies the bifurcation conditions. But this choice is somewhat arbitrary (and could be influenced by numerical details). A different choice would result in a slightly different boundary definition. However, because the boundaries are only used for coarse-graining, this is not a problem: small differences will only affect the negligible probabilities located in the transition regions (30, 41) between macrostates and will not affect thermodynamic properties. For example, the bifurcation illustrated in Fig. 1 happened to use the Thr-2 C_γ–Thr-7 C_γ distance, but there would not have been a significant difference in the macrostate boundary (when projected onto the internal coordinate space) if the Thr-2 C_β–Thr-7 C_β distance had been used instead. Because the fluctuations in $r_{Thr2C_\gamma - Thr7C_\gamma}$ and $r_{Thr2C_\beta - Thr7C_\beta}$ are correlated, the effective-energy barriers in both distance variables are simultaneously removed when the parental macrostate is subdivided by using either distance variable.

Multiple recursive dissections as T was lowered yielded macrostate specifications that were products of inequality constraints involving multiple distance variables (see Eq. **3**). Computational sampling of the macrostates converged rapidly at all T because the bifurcation procedure ensured that there were never any significant barriers to sampling within a single macrostate.

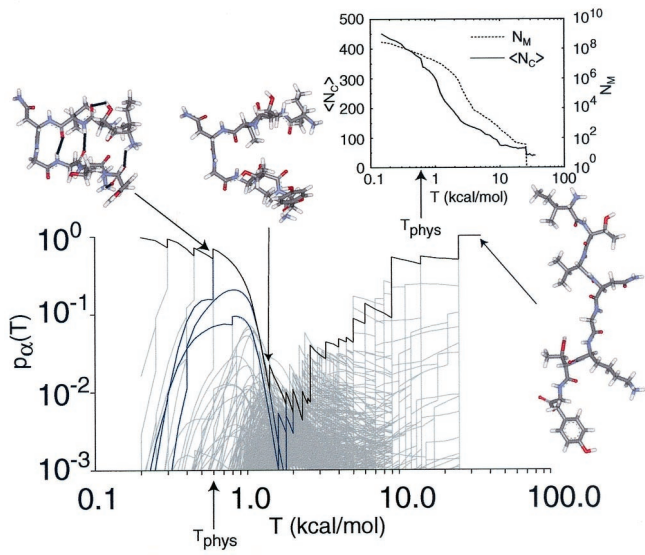

Fig. 3. Probability macrostate tree for BH8. The trajectories leading to PRSM members are highlighted. The peptide figures show the average macrostate conformations at three different temperatures on the trajectory that leads to the most probable macrostate (i.e., the native state). The PRSM trajectories are darkened. The inset plots the T-dependent number of macrostates $N_M(T)$ (Eq. **15**) and the probability-weighted mean number of nonlocal atom pairs in contact $\langle N_C\rangle(T)$ (Eq. **14**).

Properties of the Macrostate Probability Tree. We first analyzed the hierarchical organization of the macrostates by plotting the macrostate probabilities $p_\alpha(T)$ as a function of T (Fig. 3). Each continuous line segment is a macrostate branch, and each macrostate bifurcation corresponds to a fork. This tree provides significant insight into the underlying structure of the potential energy landscape and illustrates some features that will be common in all cases: (*i*) At high T (here, >25 kcal/mol) a single

BIOPHYSICS

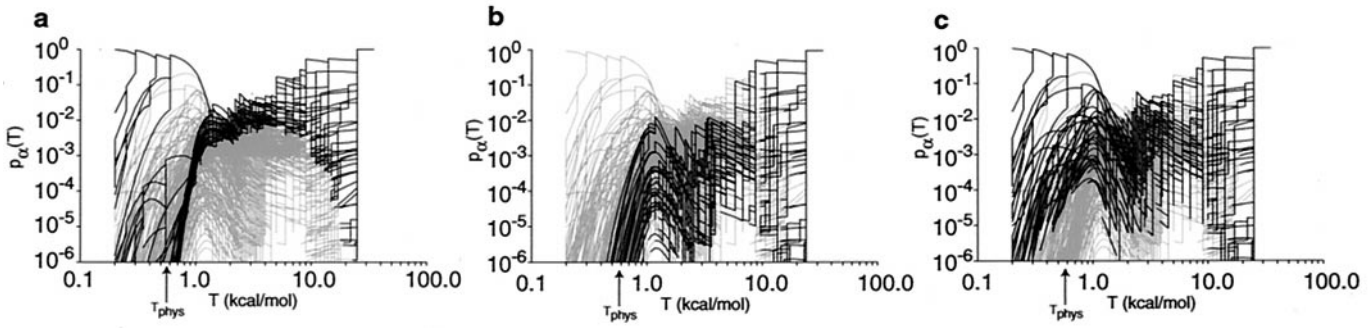

Fig. 4. Branch-selection strategies. The BH8 macrostate tree shown in Fig. 3 is replotted in gray. The subsets of 30 trajectories that were followed by a trajectory-selection algorithm using free energy $F_\alpha(T)$ (*a*), mean energy $E_\alpha(T)$ (*b*), and exergy $\Xi_\alpha(T)$ (*c*) are darkened. The PRSM corresponds to the four trajectories with lowest Ξ_α (equivalently, lowest F_α) states at T_{phys}.

macrostate contains all the ensemble probability. (*ii*) The number of macrostates, $N_M(T)$, computed by integrating Eq. **15**, increases geometrically with decreasing T (see *Inset*, Fig. 3), and probability is distributed between many macrostates in the temperature midrange. (*iii*) $N_M(T)$ continues to increase with decreasing T until, as $T \to 0$, each steepest-descent catchment region corresponds to a macrostate. (*iv*) As $T \to 0$, the macrostate that contains the energy global minimum (whose trajectory is the black line in Fig. 3) captures all the probability. This will not necessarily correspond to the macrostate that contains the most probability at T_{phys} (i.e., the folded state, if there is one). (*v*) The PRSM is not huge for peptides and proteins that assume a folded state. For example, 93% of the BH8 probability at T_{phys} is contained in only four macrostates.

The nonuniform variation of N_M with T (Fig. 3 *Inset*) suggests that the structure of the BH8 potential energy landscape can be qualitatively classified into four different temperature regimes. Analyzing the bifurcations in these regimes indicates that: (*i*) The burst of bifurcations at $T \sim 25$ kcal/mol is associated with energy barriers imposed by the rigid covalent geometry used by ECEPP/3. These only affect the local structure of the molecule. (*ii*) N_M increases only slowly until $T \sim 4$ kcal/mol where attractive forces become strong enough to initiate collapse with a consequent increase in the mean number of atom pairs in contact, $\langle N_C \rangle$ (see Eq. **14**). Below this "transition temperature" the van der Waals attraction becomes significant and probability can get trapped in an increasing number of dynamic catchment regions. (*iii*) The rate of increase decreases for $T \sim 1.5$ kcal/mol, probably because most van der Waals contacts have already been made.

Branch-Selection Algorithms. All branches having $p_\alpha \geq 10^{-3}$ are plotted in Fig. 3 for reference. But as discussed above, our goal is to determine whether the PRSM can be found while computing only a small fraction of the branches selected during the annealing process. The BH8 macrostate tree provides an illustrative example: Because the PRSM members have the lowest $F_\alpha(T_{phys})$, it was plausible *a priori* that PRSM ancestors might be identified at high T as the states of lowest $F_\alpha(T)$. But Fig. 3 shows that, for $1.4 < T < 2$ kcal/mol, some of the PRSM trajectories pass through temperature regions where they have very small probability (i.e., high F_α). Therefore, they would not be followed by a low-F_α branch-selection algorithm (i.e., which kept only a fixed number of the lowest F_α macrostates at each T) unless a large number of trajectories were followed.

We examined two additional branch selection strategies, branch selection based on: (*i*) low E_α (low energy), which ignores entropy in making predictions, and (*ii*) low Ξ_α (low exergy, $\Xi_\alpha \equiv E_\alpha - T_{phys}S_\alpha$; ref. 42), which gives some weight to entropy, but not as much as does the low-F_α strategy. To compare the predictive power of these three strategies, we determined the trajectory subsets that emerged when only 30 trajectories were followed by using either F_α, E_α, or Ξ_α as the branch-selection criterion (dark trajectories in Fig. 4 *Left*, *Middle*, and *Right*, respectively). In this test, success is measured by the number of high-probability (at T_{phys}) macrostates that are tracked by each branch-selection method. Clearly, the exergetic selection has a much stronger propensity to track macrostates that have low free energy at T_{phys}; thus it provides a superior branch-selection algorithm to both energy and free energy. Similar analysis of the pentapeptide Met-enkephalin showed that Ξ_α is a good predictor for this case as well (data not shown).

Next Steps. This study provides proof of principle for top-down free-energy minimization, but efficiency will have to be increased for much larger problems. Many improvements are possible. For example, instead of monitoring effective-energies for each of the distance variables, because distance variable redundancy grows with the number of atoms, for large systems it should be sufficient to consider only a representative subset of distance variables that includes an appropriately selected mix of small and large distances. For example, including all atom pairs that were separated by 1, 2, 4, . . . covalent bonds would result in a representative subset having only $O(N \log N)$ distance variables, where N is the number of atoms. It also may be possible to use smoothing methods (33) to more rapidly approximate branches at higher temperatures, and to combine these with principal-component and -coordinate methods (13, 17) to eliminate inessential degrees of freedom. And the accuracy (and cost) of numerical integrations at $T > T_{phys}$ can be adaptively relaxed consistent only with the need to detect bifurcations and apply the branch-selection criteria.

The most significant efficiency increases probably will come from improved branch-selection strategies: The ones tested here simply compare macrostate properties at the same T, pick those of highest rank, and discard the others. But this all-or-nothing approach tends to concentrate excessively on closely related macrostates and does not allow for reexamination of previously discarded macrostates. It should be possible for a more sophisticated algorithm to probabilistically allocate computational effort and to dynamically adjust the balance between depth and breadth of search while simultaneously searching macrostates to different depths in T. In addition, thermodynamic branch-selection parameters can be augmented with database-derived empirical parameters such as Ramachandran and secondary structure propensities that might be more powerful at high T. It also will be interesting to explore the possibility of making the potentials themselves T-dependent to improve branch selection without affecting the T_{phys} behavior.

Potentials with empirical solvation, such as used here, have

been successful in predicting the structures of peptides containing up to about 60 amino acids (43). And they are currently useful for perturbative folding problems such as refining experimentally-determined structures or homology modeling predictions. Although we have used an *ab initio* problem as an example, the top-down approach also can be used for perturbative folding. As with all potential energy-based methods, the accuracy of top-down searches ultimately will depend on the development of improved potentials. The hierarchical approach can assist this development by providing a such developments in two ways. In addition to helping perform the global minimization required to determine the predictions of a potential function, macrostate trees may help by providing a meaningful measure of the distance between the experimental and potential-predicted macrostates: The temperature at which ancestors first diverge (i.e., analogous to the time of evolutionary divergence) may be more valuable than conventional rms deviation measures for systematically improving potential energy performance.

Although not the main focus here, we note that like hierarchical methods based on grouping local minima macrostate analysis provides the information needed for a thermodynamic analysis of ensembles and for an approximate master equation description of the dynamics of folding and conformational change. It has the advantage that branch selection can be used to restrict computational effort to the branches of physical relevance (i.e., significant probability at T_{phys}). Moreover, in the top-down approach there is no need to separately search for saddle points and reaction coordinates between all pairs of catchment regions because the bifurcation temperatures estimate the barrier activation free energies and if more accurate results are needed, the bifurcating distance variables and (the negative logarithms of) their probability histograms provide reaction coordinates and effective-potential functions, respectively, for computing isothermal folding at T_{phys}. By projecting the kinetic description in the macrostate basis onto a reduced representation by using order parameters such as density and number of native contacts, it should be possible to examine the dynamics predicted by the macrostate tree for funneling and related properties (3, 8).

It is possible that the hierarchical inheritance property of exergy that makes it a useful branch-selection parameter is a general statistical consequence of the fact that the peptide potential energy function is a sum of a large number of semi-independent terms; or this may depend on other specific properties. Nymeyer *et al.* (44), comparing three lattice model systems, recently showed that the number of native contacts was a useful order parameter for dynamical folding-rate calculations only when the potential energy landscape possessed funneling properties defined by the relationship between the glass and folding temperatures. It seems plausible that funneling properties would favor accurate branch-selection, though it is not evident that they are required for it. Hierarchical analysis of many systems will be needed to study this and to determine whether there are prerequisites, beyond semiseparability of the potential, that are required for effective branch selection.

In summary, top-down macrostate tree analysis provides a potentially advantageous alternative to local minimum-based approaches for analyzing the hierarchical structure of protein energy landscapes. It naturally includes entropic as well as energetic effects and can identify and exploit hidden hierarchical properties in new types of global search procedures. Developing further improved branch-selection algorithms and understanding how they perform as protein size increases are important future tasks.

We thank J. Gans for countless valuable discussions, S. Berry for a critical review of the manuscript, R. Elber, J. Straub, D. Thirumalai, and P. Wolynes for comments on the manuscript, and National Science Foundation Grant CCR-9988519 and the Intel Corp. for financial support.

1. Bryngelson, J. D. & Wolynes, P. G. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 7524–7528.
2. Boczko, E. M. & Brooks, C. L. (1995) *Science* **269,** 393–396.
3. Wolynes, P. G., Onuchic, J. N. & Thirumalai, D. (1995) *Science* **267,** 1619–1620.
4. Shakhnovich, E. I. (1997) *Curr. Opin. Struct. Biol.* **7,** 29–40.
5. Plotkin, S. S., Wang, J. & Wolynes, P. G. (1996) *Phys. Rev. E* **53,** 6271–6296.
6. Plotkin, S. S., Wang, J. & Wolynes, P. G. (1997) *J. Chem. Phys.* **106,** 2932–2948.
7. Plotkin, S. S., Wang, J. & Wolynes, P. G. (1997) *Physica D* **107,** 322–325.
8. Onuchic, J. N., Luthey-Schulten, Z. & Wolynes, P. G. (1997) *Annu. Rev. Phys. Chem.* **48,** 545–600.
9. Stillinger, F. H. & Weber, T. A. (1984) *Science* **225,** 983–989.
10. Noguti, T. & Gō, N. (1989) *Proteins* **5,** 97–103.
11. Troyer, J. M. & Cohen, F. E. (1995) *Proteins* **23,** 97–110.
12. Daura, X., van Gunsteren, V. F. & Mark, A. (1999) *Proteins Struct. Funct. Genet.* **34,** 269–280.
13. Elmaci, N. & Berry, R. S. (1999) *J. Chem. Phys.* **110,** 10606–10622.
14. Czerminski, R. & Elber, R. (1990) *J. Chem. Phys.* **92,** 5580–5601.
15. Becker, O. M. & Karplus, M. (1997) *J. Chem. Phys.* **106,** 1495–1517.
16. Becker, O. M. (1997) *Proteins Struct. Funct. Genet.* **27,** 213–226.
17. Becker, O. M. (1998) *J. Comp. Chem.* **19,** 1255–1267.
18. Wales, D. J., Miller, M. A. & Walsh, T. R. (1998) *Nature (London)* **394,** 758–760.
19. Doye, J. P., Miller, M. A. & Wales, D. J. (1999) *J. Chem. Phys.* **110,** 6896–6906.
20. Miller, M. A. & Wales, D. J. (1999) *J. Chem. Phys.* **111,** 6610–6616.
21. Schön, J., Putz, H. & Jansen, M. (1996) *J. Phys. Condens. Matter* **8,** 143–156.
22. Schön, J. (1996) *Ber. Bunsenges.* **100,** 1388–1391.
23. Schön, J. & Sibani, P. (2000) *Europhys. Lett.* **49,** 196–202.
24. Orešič, M. & Shalloway, D. (1994) *J. Chem. Phys.* **101,** 9844–9857.
25. Church, B. W., Orešič, M. & Shalloway, D. (1996) *Tracking Metastable States to Free-Energy Global Minima,* DIMACS Series in Discrete Mathematics and Theoretical Computer Science, eds. Pardalos, P., Shalloway, D. & Xue, G. (Am. Math. Soc., Providence, RI), Vol. 23, pp. 41–64.
26. Church, B. W., Ulitsky, A. & Shalloway, D. (1999) *Adv. Chem. Phys.* **105,** 273–210.
27. Ball, K. D. & Berry, R. S. (1998) *J. Chem. Phys.* **109,** 8541–8556.
28. Ball, K. D. & Berry, R. S. (1998) *J. Chem. Phys.* **109,** 8557–8572.
29. Stillinger, F. H. & Weber, T. A. (1982) *Phys. Rev. A* **25,** 978–989.
30. Shalloway, D. (1996) *J. Chem. Phys.* **105,** 9986–10007.
31. Sherrington, D. (1997) *Physica D* **107,** 117–121.
32. Straub, J. E. & Thirumalai, D. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 809–813.
33. Shalloway, D. (1997) in *Variable-Scale Coarse-Graining in Macromolecular Global Optimization,* eds. Biegler, L., Coleman, T., Conn, A. R. & Santosa, F. (Springer, New York), pp. 135–161.
34. Gō, N. & Scheraga, H. A. (1976) *Macromolecules* **9,** 535–542.
35. Vanderbilt, D. & Louie, S. G. (1984) *J. Comp. Phys.* **56,** 259–271.
36. Church, B. W. & Shalloway, D. (1996) *Polymer* **37,** 1805–1813.
37. Frantz, D. D., Freeman, D. L. & Doll, J. D. (1990) *J. Chem. Phys.* **93,** 2769–2784.
38. Ramirez-Alvarado, M., Blanco, F. J. & Serrano, L. (1996) *Nat. Struct. Biol.* **3,** 604–611.
39. Némethy, G., Gibson, K. D., Palmer, K. A., Yoon, C. N., Paterlini, G., Zagari, A., Rumsey, S. & Scheraga, H. A. (1992) *J. Phys. Chem.* **96,** 6472–6484.
40. Abagyan, R. (1997) in *Simulation of Biomolecular Systems: Theoretical and Experimental Applications,* eds. van Gunsteren, W. F., Weiner, P. K. & Wilkinson, A. J. (ESCOM, Leiden, The Netherlands), pp. 363–394.
41. Ulitsky, A. & Shalloway, D. (1998) *J. Chem. Phys.* **109,** 1670–1686.
42. Bejan, A. (1988) *Advanced Engineering Thermodynamics* (Wiley, New York), pp. 111–145.
43. Pillardy, J., Czaplewski, C., Liwo, A., Lee, J., Ripoll, D. R., Kazmierkiewicz, R., Oldziej, S., Wedemeyer, W. J., Gibson, K. D., Arnautova, Y. A., *et al.* (2001) *Proc. Natl. Acad. Sci. USA* **98,** 2329–2333. (First Published February 20, 2001, 10.1073/pnas.041609598)
44. Nymeyer, H., Socci, N. D. & Onuchic, J. N. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 634–639.